Phishing Website Detection Using Machine Learning Techniques

Prerna Garsole
Software Engineering Department
San Jose State University
California, USA
prerna.garsole@sjsu.edu

***Abstract*—In this research, I am addressing the issue of phishing website identification in this research by utilizing the most recent machine-learning techniques. Think of the internet as a vast landscape in which our task is to find phishing sites, or websites that try to trick users into providing sensitive information. After that, I retrieved 17 different attributes to aid in the interpretation of these URLs. These traits allow to identify the distinctive qualities of each website. Other features concentrate on the domain, others on the address bar, and some on HTML and JavaScript. Right now, the interesting element is machine-learning models is several models have been trained to differentiate between phishing and legitimate websites, including support vector machines, XG-Boost, The Random Forest, Auto-encoder with Neural-Networks, and Multilayer-Perceptron's. I have trained this group of models using our dataset. It's similar to teaching them the finer points of a language that only they understand—the language of URLs. I then gave them a new set of URLs to test what they had learned. These results helped to determine which models are most effective at identifying phishing attempts. The objective is to create a system that stays one step ahead of all the malicious actors by using state-of-the-art algorithms and understanding the traits that indicate a dubious website. As a result, the research's practical consequences improved the security of your online interactions by increasing the accuracy of identifying phishing websites.**

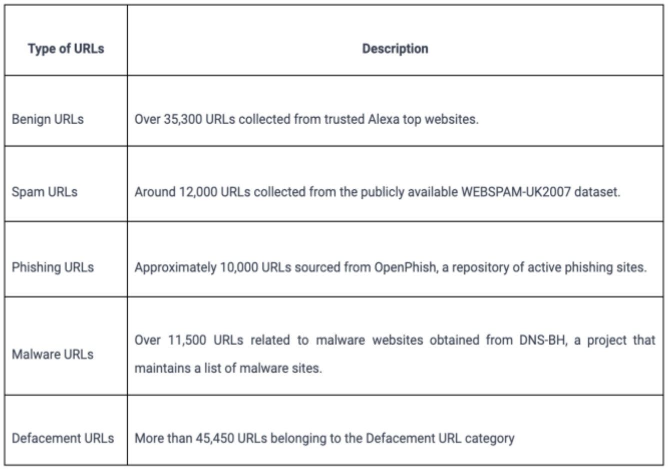
***Index Terms*—Phishing, XG-Boost, Random-Forest, Decision-Trees**

I. INTRODUCTION

In today's connected and useful digital world, phishing is a real danger that we cannot at all ignore. Cybercriminals play a game where they trick people into providing sensitive information, such as login passwords or private financial information. Both individuals and businesses are impacted by financial loss and identity theft. The study contributes significantly to the global security of digital authentication by examining the state-of-the-art in phishing detection. By reviewing earlier research, I built on an extensive array of methods and concepts related to phishing detection. Machine learning, which is changing the game, looks for patterns that unmistakably point to "phishing!" like a digital detective. It is more than just understanding the current situation; it entails actively shaping the future. This exciting journey combines technological innovation with human behavior to investigate the relationship between machine learning and phishing detection. The goal is to protect the integrity of our digital lives in this highly interconnected world by offering solutions and thought-provoking nuggets of information that elevate the current cybersecurity discussion.

II. PHISHING DATASET

Valid URLs are gathered from the University of New Brunswick's dataset, which is available at https://www.unb.ca/cic/datasets/url-2016.html.



Type of URLs	Description
Benign URLs	Over 35,300 URLs collected from trusted Alexa top websites.
Spam URLs	Around 12,000 URLs collected from the publicly available WEBSPAM-UK2007 dataset.
Phishing URLs	Approximately 10,000 URLs sourced from OpenPhish, a repository of active phishing sites.
Malware URLs	Over 11,500 URLs related to malware websites obtained from DNS-BH, a project that maintains a list of malware sites.
Defacement URLs	More than 45,450 URLs belonging to the Defacement URL category

Table. 1: Dataset Details

5000 URLs are chosen at random from the list. The collection focuses on URLs that are commonly used as a gateway for illicit activity conducted online. Its purpose is to help identify and classify dangerous URLs according to the sort of threat they pose. I have processed the dataset by with various steps. These stages includes performing ETL that is extract transforming and loading the data after cleaning of the dataset.

III. METHODOLOGY AND EXPERIMENT

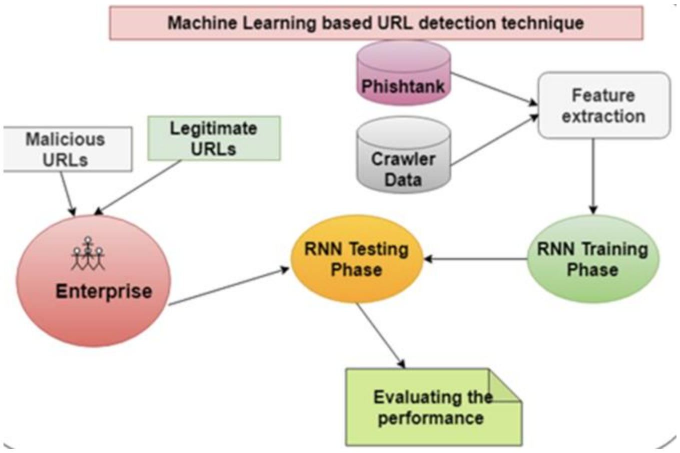


FIGURE 1: ML models Methodology

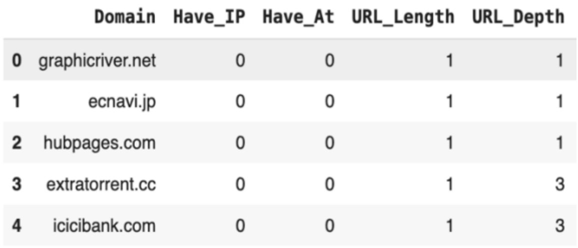
I have used the methodical and effective approach that included extracting and analyzing the different 17 important information from URLs in order to detect phishing websites. Address bar features, domain-based features, and HTML/JavaScript-based features were the three main primary categories into which these features were divided.

1) *Features of the Address Bar*

I have examined and found the URL's address bar, taking note of variables such as the URL's domain, whether it contains the '//' symbol, which indicates redirection, whether an IP address appears, whether the domain name contains the characters 'http' or 'https,' whether the '@' symbol is present, whether URL shortening services are used, and how long the URL is.

2) *Features of the domains*

I have also additionally looked at the attributes related to the domain itself. I took into account DNS records, domain age, website traffic information, and domain expiration date.



	Domain	Have_IP	Have_At	URL_Length	URL_Depth
0	graphicriver.net	0	0	1	1
1	ecnavi.jp	0	0	1	1
2	hubpages.com	0	0	1	1
3	extratorrent.cc	0	0	1	3
4	icicibank.com	0	0	1	3

Table. 2: Dataset Processing

3) *JavaScript and HTML-based Qualities*

I have also examined the way the website was built, evaluating things like frame redirection, right-click disablement, status bar and customization, website forwarding, the use of '-' as a prefix or suffix in the domain name, and the depth of URL

IV. MACHINE LEARNING MODELS AND TRAINING

The data set comes under the classification problem, as the input URL is classified as phishing (1) or legitimate (0). The supervised machine learning models (classification) considered to train the dataset are as follows.

1) Decision-Tree

In comparison to more advanced models, this Tree model may show limits when it comes to processing complicated relationships within the data, despite being a vital weapon in the arsenal for phishing website identification.

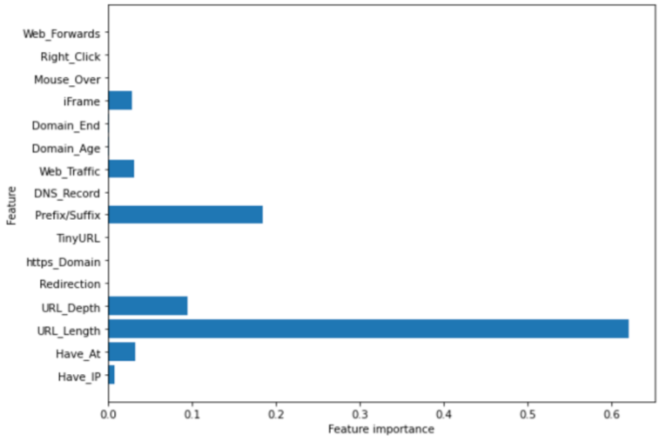


FIGURE 2: Result1

2) Random-Forest

This model not only performed exceptionally well, but it also showed remarkable efficacy in using the selected criteria to distinguish between reputable and phishing websites. The random forest model is used in time series data. The model has less accuracy in testing the phishing website. This model has recognized the phishing sites as listed in the results section of the project experimentation. Please refer table 3 of comparison

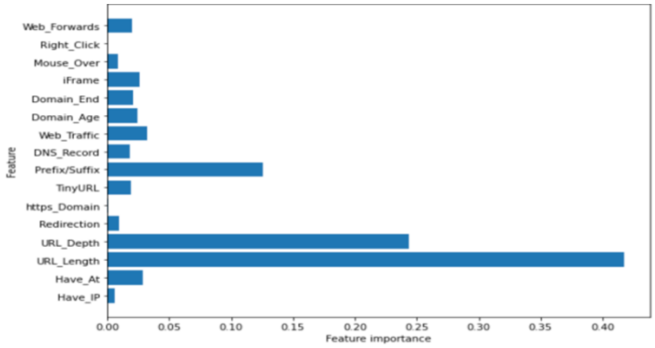


FIGURE 2: Result2

3) Multilayer-Perceptron's

This showed remarkable capacity, using complex neural networks (NNs) with neural topologies to help correctly identify phishing websites in the dataset.

4) XG-Boost

Among machine-learning models, XG-Boost performed very well and emerged as the front-runner. It did this by skillfully utilizing its boosting strategy to improve accuracy in differentiating between phishing and legitimate websites.

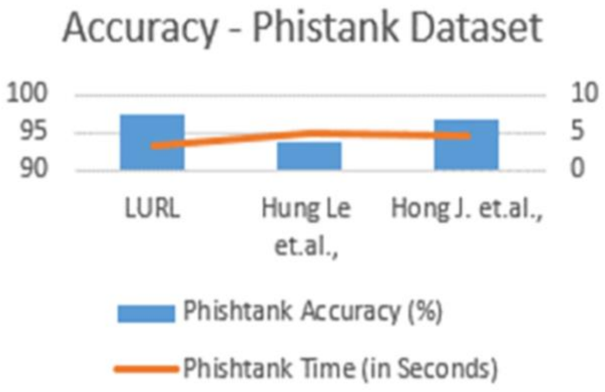


FIGURE 3: Model Accuracy

5) Autoencoder-Neural-Network

This model shown remarkable potential in deriving meaningful representations from the dataset, which might help identify phishing attempts, thanks to its unsupervised learning capability.

6) Support-Vector-Machines

Using their capacity to establish the best possible decision boundaries to distinguish between phishing and trustworthy websites in the dataset under analysis. This model is a type of supervised learning model that is used to analyzed the data for classification. Hence this model is used to classify the website as phishing website or as legitimate website along with its comparison with other models.

V. RESULTS AND ANALYSIS

1. The XG-Boost's Better Performance

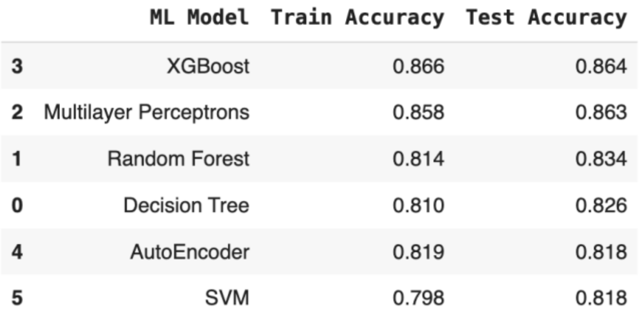
When it comes to distinguishing between phishing and trustworthy websites over the internet, XG-Boost has shown to be the most remarkably effective. It is exceptional results in the test and training datasets demonstrate how resilient it is when picking up knowledge and making generalizations from the given characteristics.

2. The Effectiveness of Features

A key factor in the improving the detection accuracy of the phishing websites was the carefully selected collection of features that were derived from the address bar, and domain-based, and HTML/JavaScript-based attributes. This indicates that a rich and educational dataset was made available for model training through the combination of these properties.

VI. MODELS COMPARISON

A data frame is constructed in order to compare the models' performances. The lists made to hold the model's findings are the columns in this data frame.



	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.866	0.864
2	Multilayer Perceptrons	0.858	0.863
1	Random Forest	0.814	0.834
0	Decision Tree	0.810	0.826
4	AutoEncoder	0.819	0.818
5	SVM	0.798	0.818

Table. 3: Comparison

From the above table it shows that the best model performed is XG Boost model. That has the highest train accuracy as 0.866 and test accuracy as 0.864. The second highest model is the multilayer model that has the 2nd highest accuracy in test and training which is 0.858 and 0.836 respectively.

The models are again compared with another followed by decision model and support-vector which are the supervised learning models. These are used to test and analyze the phishing websites. The Decision model has the precision as 0.810 and 0.826 as the training and test measures of the accuracy.

The lowest train accuracy is for SVM which is 0.798 and the test accuracy is 0.818 which the lowest accuracy for testing and detecting the phishing websites. These measures and metrics for phishing website detection are directed in above table with each model's accuracy of training and test accuracy.

X. CONCLUSION

During the evaluation, XG-Boost demonstrates superior performance in both of the training and test datasets. The chosen features are effectively enhancing the detection of phishing websites. The higher side of the XG-Boost model justifies its retention for future use.

This robust and efficient solution significantly aids in accurately identifying malicious URLs. The incorporation of well-defined features and advanced machine learning techniques, particularly XG-Boost, contributes to improved accuracy and efficiency in phishing website detection.

PROJECT MATERIALS

GitHub link - https://github.com/prernagarsole/Phishing-Website-Detection-by-Machine-Learning-Techniques

Dataset files - https://github.com/prernagarsole/Phishing-Website-Detection-by-Machine-Learning-Techniques/tree/main/DataFiles

Processed and cleaned data after performing ETL - https://github.com/prernagarsole/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/main/DataFiles/5.urldata.csv

Presentation Link - https://github.com/prernagarsole/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/main/Presentation.pptx

URL Extraction Files - https://github.com/prernagarsole/Phishing-Website-Detection-by-Machine-Learning-Techniques/tree/main/URL%20Extraction%20files

Train_models_Phishing_Website_Detection Files - https://github.com/prernagarsole/Phishing-Website-Detection-by-Machine-Learning-Techniques/tree/main/Train_models_Phishing_Website_Detection

REFERENCES

[1] S. Leung, K. Chan, F. Chung, and G. Ngai, "A probabilistic rating inference framework for mining user preferences from reviews," World Wide Web, vol. 14, no. 2, pp. 187-215, 2011. [Online]. Available: doi: 10.1007/s11280-011-0117-5. M.

[2] Aljawarneh, "Phishing websites detection using machine learning: A systematic review," Journal of King Saud University - Computer and Information Sciences, 2020. [Online]. Available: doi: 10.1016/j.jksuci.2020.11.034.

[3] R. A. Alshammari, H. Alshammari, and A. S. Al-Daraiseh, "Machine learning-based phishing detection: A review," Journal of Information Security and Applications, vol. 50, 2020. [Online]. Available: doi: 10.1016/j.jisa.2019.102408.

[4] UNB CIC Datasets. [Online]. Available: https://www.unb.ca/cic/datasets/url-2016.html

[5] PhishTank Developer Information. [Online]. Available: https://www.phishtank.com/developer_info.php

[6] "Phishing Website Detection Using Machine Learning Techniques," [Online]. Available: https://archive.ics.uci.edu/dataset/327/phishing+websites